Omni-ID: Holistic Identity Representation Designed for Generative Tasks — Appendix —

Guocheng Qian Kuan-Chieh Wang Or Patashnik Negin Heravi
Daniil Ostashev Sergey Tulyakov Daniel Cohen-Or Kfir Aberman
Snap Research
https://guochengqian.github.io/project/

A. Experiment Details

A.1. Baseline Details

CLIP. Throughout our experiments, we use CLIP-H from OpenAI [11] as the feature extractor, which works slightly better than CLIP-B/L. We use the full representation, *i.e.* 257 tokens (256 spatial tokens with 1 class token) from the second last layer following IP-Adapter Full [16], which improves ID preservation compared to using only class token or a reduced number of tokens (*e.g.* 16). For multiple inputs, we use token concatenation following IP-Adapter, which outperforms simple averaging. For fair comparisons, we train IP-Adapter with CLIP representation in the same Flow Matching Decoder stage for 5K steps with an effective batch size 32 (roughly 1 epoch in MFHQ). The convergence happens at around 4K steps.

ArcFace. We use the ArcFace [4] model from insightface [3] throughout the experiments. We project Arcface embedding from $\mathbb{R}^{1\times512}$ to 256 tokens with 1280 channels $\mathbb{R}^{256\times1280}$, which is comparable to CLIP and Omni-ID in terms of representation size. Using 256 tokens improves its ID preservation compared to using 4 or 16 tokens only in IP-Adapter FaceID [16], and outperforms other reduced number of tokens (64). We concatenate representations in the token dimension for multiple inputs, where averaging merging also reaches a similar results for ArcFace representation. We train IP-Adapter with ArcFace representation in the Flow Matching Decoder stage by 75K steps to converge. Compared to CLIP and Omni-ID which take about 5K steps, the convergence of ArcFace is rather slow, due to its over-compactness for generative tasks.

ArcFace+CLIP. Following IP-Adapter FaceIDPlus [16], Arc-Face+CLIP baseline projects ArcFace tokens from the average ArcFace embeddings in $\mathbb{R}^{1\times512}$ to 256 queries in $\mathbb{R}^{256\times1280}$, where each individual CLIP features in $\mathbb{R}^{257\times1280}$ are used as keys and values to aggregate multiple inputs. For a fair comparison to Omni-ID, the same transformer with self attention layers is used to merge features. Both the improved number of tokens and the self-attention layers in transformer improves the face quality compared to the original implementation in FaceIDPlus, where 4 or 16 tokens are used as query and Q-former [7] without self attentions are employed.

A.2. Decoders and Training Details

Decoders details.

- *Masked transformer decoder.* MTD is built by 6 CA blocks and 2 SA blocks, which reaches high-quality reconstruction while smaller number of decoder blocks might compensate encoder quality due to the lower decoding ability. Mask ratio of MTD is set to 95%, *i.e.* 5% patches are visible during training, which leads better encoder performance in downstream tasks than mask ratio 85% or 99%. The patch size for the decoder is set to 14×14 to balance the speed and quality.

- *Flow-matching denoising decoder.* For the Flow-Matching Decoder, FLUX dev [1] serves as the base model. We implement a FLUX-based version of IP-Adapter [16], where the Omni-ID representation is injected into all blocks, including MM-DiT and DiT blocks, via learnable decoupled attention layers. Injecting into both block types results in slightly better quality compared to injecting into only MM-DiT blocks or only DiT blocks, although this improvement is not critical. Each decoupled attention layer optimizes a single linear projection to map ℓ from $\mathbb{R}^{L\times C}$ to $\mathbb{R}^{L\times3250}$, where 3250 is the channel size used in FLUX. During the Flow-Matching Decoder stage, the Omni-ID encoder and the projection layers of the decoupled attention layers are optimized, while the original parameters in FLUX remain frozen.

Training Details. Omni-ID uses a two-stage few-to-many identity reconstruction training process: the MTD stage and the Flow Matching Decoder stage. The MTD stage is trained on our MFHQ dataset at an image resolution of 448 using a constant learning rate of $1e^{-4}$, an effective batch

Figure I. **Gallery of Omni-ID in personalized T2I generation.** Omni-ID enables high identity preservation. Results achieved by injecting Omni-ID representation through IP-Adapter [16] into the frozen FLUX dev model [1] without LoRA [6].

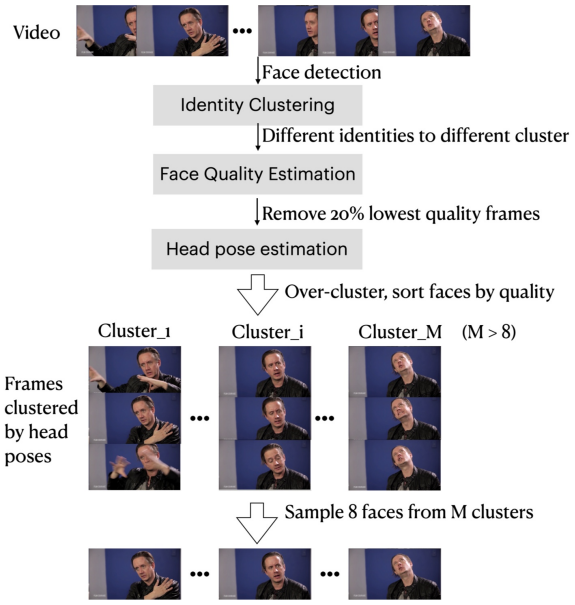

Figure II. **Illustration of MFHQ Creation**. Given a video, we first detect faces and distribute identities to different clips by a threshold based on the cosine distance of face embeddings [4]. Then, a face quality estimation [2] is applied to sort the quality of frames within each identity. 20% faces with lowest quality are removed. A head pose estimation [13] is employed to estimate the poses for each face which are used to cluster the frames into $M = 16$ clusters. Finally, 8 frames are sampled M clusters, where each cluster is only sampled at most once. The sum of absolute pose differences is assured larger than 15 degree for each pair.

size of 256 (distributed as 32 batches across 8 NVIDIA A100 GPUs), and the AdamW optimizer for 250K iterations. The Flow Matching Decoder stage is trained on the same dataset at a resolution of 512, with a constant learning rate of $1e^{-4}$, an effective batch size of 32, and the AdamW optimizer for 5K iterations. In both stages, we uniformly sample a variable number of inputs (1 to 3) and generate all 8 targets for each identity.

Downstream Details.

- *Controllable face generation.* For all experiments, we freeze the face representation encoder and optimize both the ControlNet and IP-Adapter using a constant learning rate of $2e^{-5}$ and an effective batch size of 16 for 15K steps. The models are trained on MFHQ with a variable number of inputs (uniformly sampled between 1 and 7) and a single target image, all at a resolution of 512×512. All models converge well before reaching 15K steps. The ControlNet is implemented and initialized as described in [15]. The IP-Adapter is initialized from our Flow Matching Decoder. For fair comparisons, other representations (e.g., CLIP and ArcFace) also undergo the Flow Matching Decoder training stage to achieve convergence, requiring 5K steps for CLIP and 75K steps for ArcFace. In the benchmark, ground truth landmarks from the same identity are used as ControlNet inputs, and metrics are calculated between the generated images and the targets. The generation resolution is set to 512×512.

- *Personalized T2I generation.* We integrate frozen face representations into the frozen FLUX dev base model [1] using learnable decoupled attentions, following the approach outlined in IP-Adapter [16]. Injecting into MM-DiT blocks is unnecessary in personalized T2I and does not affect the image quality. The IP-Adapter is trained using a simple flow-matching loss without additional regularization (*e.g.* ID loss, alignment loss [5]) and without employing LoRA [6]. These regularization and LoRA modules are left for future study as orthogonal to our work. Our training is performed at a resolution of 512×512 for 50K steps with a constant learning rate of $1e^{-4}$, using the AdamW optimizer. Subsequently, we fine-tune the IP-Adapter at a resolution of 768×768 for 20K steps, maintaining the same hyperparameters. Models are trained on our internal purchased dataset (Getty Images). For fair comparisons, other representations, such as CLIP and ArcFace, are trained under the same hyperparameters unless otherwise noted. Due to its slower convergence compared to Omni-ID and CLIP, ArcFace requires 100K steps in the first stage to achieve convergence. Inference for this task is performed at a resolution of 1024×1024.

MFHQ Details. Refer to Fig. II how MFHQ is collected for each video clip.

B. Supplementary Experiments

B.1. Additional Controllable Face Generation

Fig. III further compares Omni-ID with ArcFace [4] and CLIP [11] in the context of controllable face generation. Unlike the benchmark case presented in the main paper, where Ground Truth landmarks were used to guide identity-specific generation, here we use the template-driven landmarks as conditions. 9 template images are collected to obtain a grid of expression and pose in FLAME code [8] through 3D mesh reconstruction by 3D landmark estimation. Then, we use the FLAME shape code for each identity with the template FLAME expression and pose code from each template to get the rigged mesh. From each mesh, 2D lanmarks are rendered as condition to generate each view at the grid for each identity.

While CLIP demonstrates strong baseline performance, it struggles with identity preservation and fails to generate realistic faces when the pose and expression differ significantly from the input images. This limitation arises because CLIP is an instance-level representation model. In contrast, our Omni-ID is an identity-level representation, specifically trained to reconstruct faces in new poses and expressions.

Table I. **Quantitative comparisons to the state-of-the-art on personalized T2I generation.** ID Similarity are computed by the cosine distance between the generated samples and the five images of each identity. We compute the average and std across identities. The base models are FLUX [1] for all methods.

Method	ID Similarity↑
IPA-FaceID	0.3535±0.1982
IPA-FaceIDPlusV2	0.4327±0.1090
IPA-Full	0.6649±0.0846
PuLID	0.7289±0.0572
IPA-Omni-ID Schnell (Ours)	0.7306±0.0793
IPA-Omni-ID (Ours)	**0.8026±0.0421**

Consequently, Omni-ID achieves significantly better identity preservation while generating new faces of the identity.

B.2. Additional Personalized Text-to-Image

Compare to State-of-the-Art. We compare Omni-ID+IP-Adapter (IPA Omni-ID) to the state-of-the-art IP-Adapter [16], InstantID [14], PhotoMakerV2 [9], PuLID [5] in Fig. IV and Fig. V when using FLUX Dev [15] and Stable Diffusion (SD) [12] as the base model, respectively. Our IPA Omni-ID trained by the simple flow matching loss without any advanced techniques such as LoRA [6], ID loss [5], aligment loss [5], stacked embedding [9], IdentityNet [14], achieves the highest ID preservation. Refer to *gallery.m4v* for all visual results of our model. Tab. I compares IPA Omni-ID with the state-of-the-art personalized T2I employed FLUX as the base model. Our IPA Omni-ID outperforms others with the highest identity similarity.

Beyond FLUX Dev Experiments. Despite Omni-ID is trained using FLUX dev [1] as the Flow Matching Decoder, Omni-ID can be applied to any other diffusion models. In this section, we use the Omni-ID encoder with IP-Adapter on FLUX Schnell [1] and SD15 [12] in the task of personalized text-to-image generation. Fig. IV and Fig. V demonstrates again the superiority of Omni-ID against other representations like CLIP and ArcFace.

References

[1] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 1, 2, 3, 4, 6

[2] Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 33:2404–2418, 2024. 3

[3] InsightFace Contributors. Insightface: 2d and 3d face analysis project. https://github.com/deepinsight/insightface, 2024. Accessed: 2024-11-15. 1

[4] Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):5962–5979, 2022. 1, 3, 5

[5] Zinan Guo, Yanze Wu, Zhuowei Chen, Lang Chen, and Qian He. Pulid: Pure and lightning ID customization via contrastive alignment. *CoRR*, abs/2404.16022, 2024. 3, 4, 6

[6] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 2, 3, 4, 6

[7] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1

[8] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 3

[9] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. *arXiv preprint arXiv:2312.04461*, 2023. 4, 7

[10] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2024. 7

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 3, 5

[12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 4, 7

[13] Nataniel Ruiz, Eunji Chong, and James M. Rehg. Fine-grained head pose estimation without keypoints. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018. 3

[14] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 4, 7

[15] XLabs-AI. Flux-controlnet collections. https://huggingface.co/XLabs-AI/flux-controlnet-collections, 2024. Accessed: 2024-11-13. 3, 4

[16] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arxiv:2308.06721*, 2023. 1, 2, 3, 4, 6

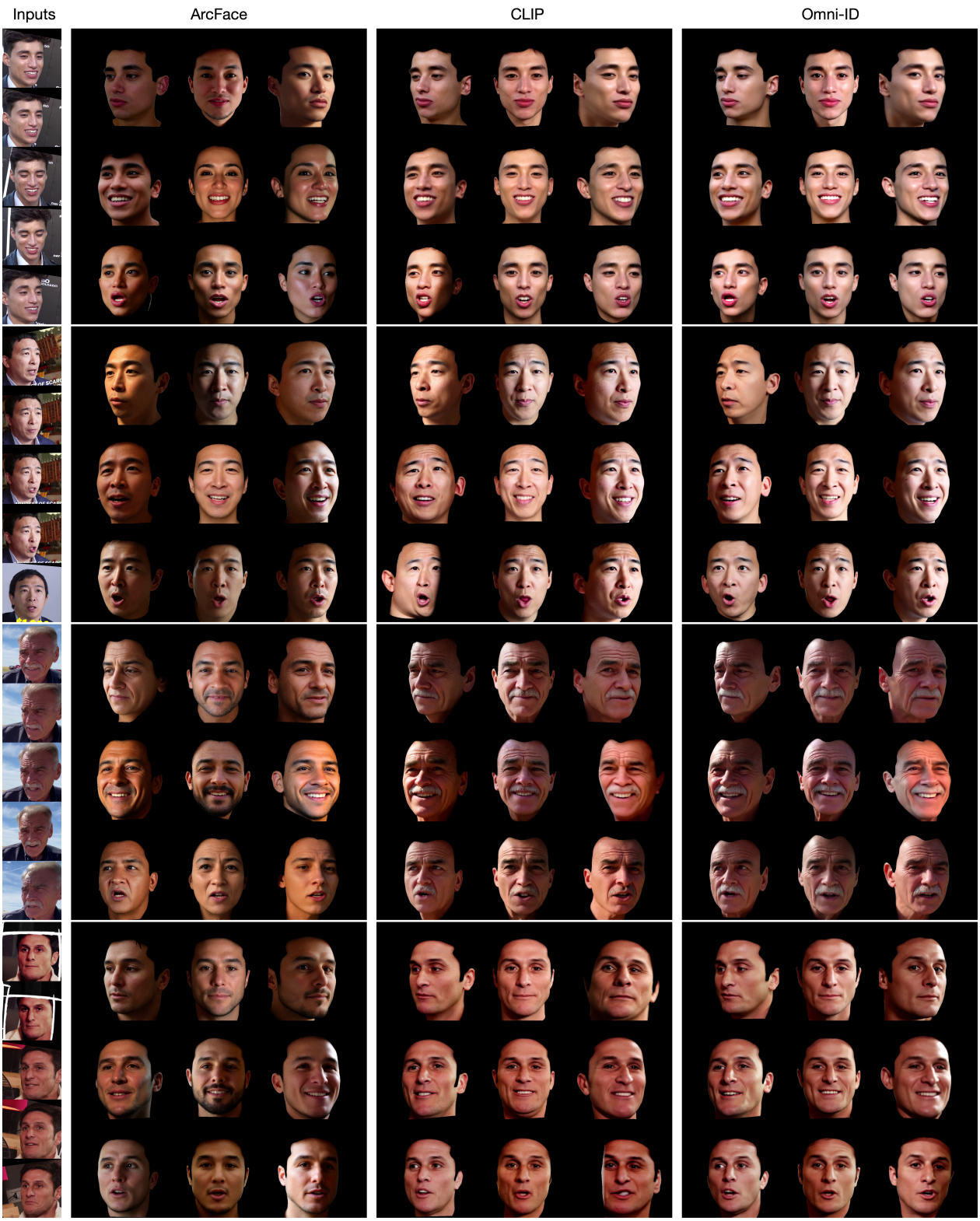

Figure III. **Qualitative comparisons to the state-of-the-art representations in controllable face generation.** We compare Omni-ID with ArcFace [4] and CLIP [11] with 5 input images. To control each face in the grid, we drive the facial landmark of each identity by the same template. Our Omni-ID achieves superior identity preservation, captures nuanced details more faithfully, and demonstrates higher adaptivity to diverse poses and expressions.

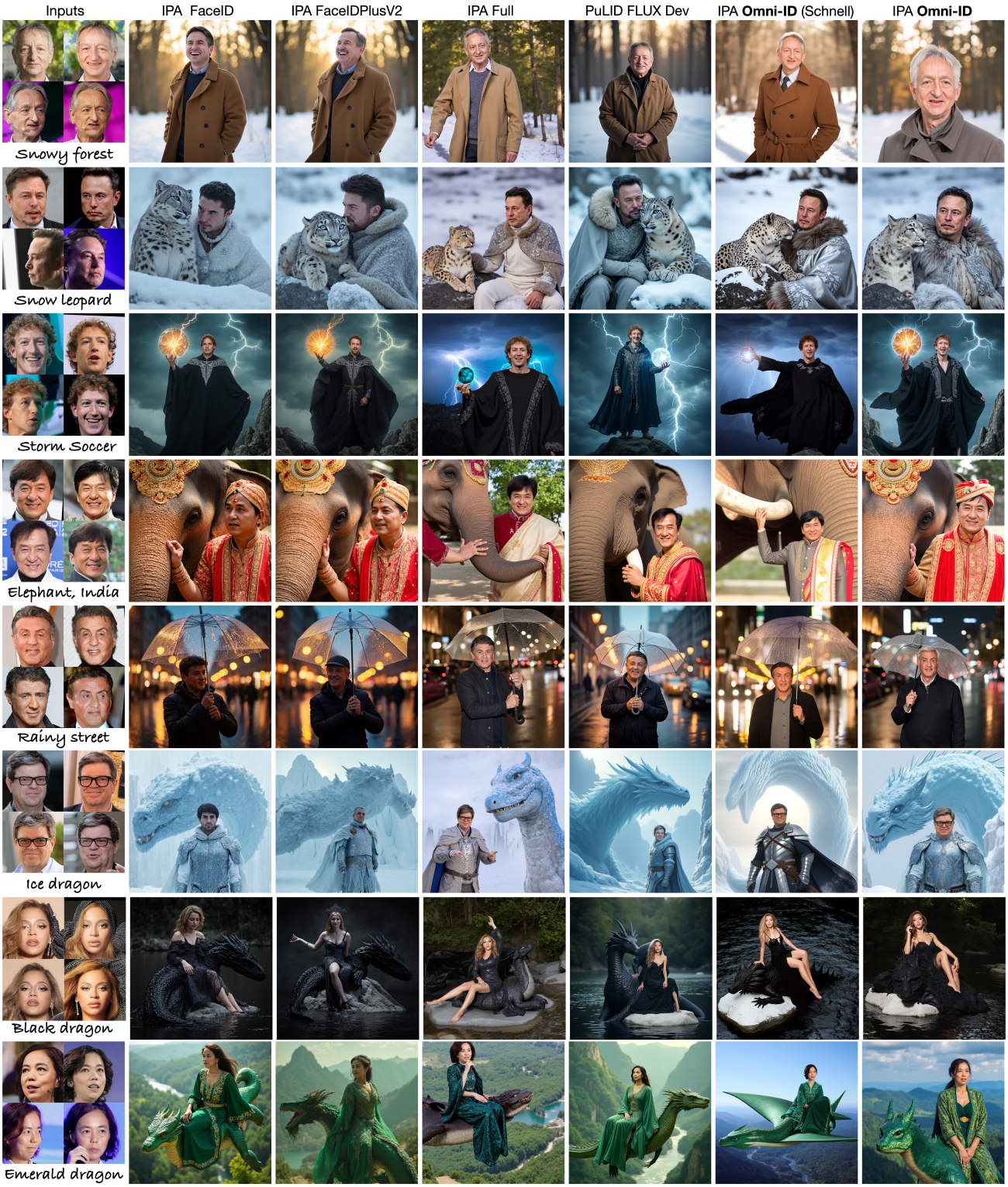

Figure IV. **Qualitative comparisons with the state-of-the-art in personalized T2I generation using FLUX [1] as the base model.** Our Omni-ID with IP-Adapter [16] without any other regularization (LoRA [6], ID loss [5], alignment loss [5]) achieves highest ID preservation. Different variants of IP-Adapter without LoRA are shown at the left side. The state-of-the-art PuLID-FLUX-v0.9.1 achieves lower face quality compared to Omni-ID. Omni-ID also works well on FLUX Schnell model, which generates each sample by 4 denoising steps.

Figure V. **Qualitative comparisons to the state-of-the-art in personalized T2I generation using Stable Diffusion [12] as the base model.** IPA-Full, IPA-Plus, and our IPA-Omni-ID use SD15 [12] as the base model, generating 512×512 resolution samples. InstantID [14] and PhotoMakerV2 [9] use SDXL [10] as the base model, generating 1024×1024 samples, which are resized to 512×512 to show with other methods side by side. Our Omni-ID with IP-Adapter without any other regularization achieves the highest ID preservation.